Log Activity

IF 4090 – Kerja Praktek

Minggu : 5

Perusahaan / Organisasi : *National Institute of Technology*, *Gifu College* (NIT – GC)

Pembimbing Kerja Praktek : Dr. Eng. Ayu Purwarianti, ST., MT.

Nama : Geraldi Dzakwan

NIM : 13514065

Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

2017

Log Activity – Minggu ke-5

Hari, Tanggal	**Kegiatan**	**Hasil**
Senin, 10 Juli 2017	1. Melakukan ekstraksi *private candidate phrase* dari kalimat yang sudah diklasifikasi *named entity tag*-nya untuk *tag* LOC (*location*), ORG (*organization*), dan PER (*person*). 2. Membuat *script* untuk megecek apakah *candidate phrase* memiliki *negative meaning*. 3. Membuat *database* untuk menyimpan *private* dan *non-private nouns/verbs* yang akan dipakai untuk menyeleksi *private candidate phrases.*	1. Fungsi yang mengekstraksi *private candidate phrases* dari *input messages* yang telah diberi *named entity tag.* 2. Python *script* untuk megecek apakah *candidate phrase* memiliki *negative meaning*. 3. MySQL *database* berisi *private* dan *non-private nouns/verbs.*
Selasa, 11 Juli 2017	1. Menambahkan mekanisme normalisasi (*stemming & lemmatization*) sesudah *pos tagging* agar memudahkan perbandingan dengan kata-kata di *database*. 2. Menyelesaikan seluruh *step* yang diperlukan untuk menyeleksi *candidate phrase* dari *tag* LOC, ORG, dan PER.	1. Fungsi untuk melakukan *normalization* sesudah dilakukan *pos tagging.* 2. Fungsi-fungsi yang menentukan apakah *candidate phrase* mengandung *private* dan/atau *non-private nouns/verbs.*
Rabu, 12 Juli 2017	1. Melakukan migrasi data *geolocation* ke MySQL database. 2. Mengimplementasi mekanisme *anonymization* untuk *geolocation* dengan menggunakan *database* tersebut. *Anonymization* dilakukan dengan mengganti tempat tersebut dengan *similar place* atau *generalized place*.	1. MySQL database berisi data *geolocation* yang dibutuhkan. 2. Fungsi untuk melakukan *anonymization* terhadap *geolocation* yang terkoneksi dengan *database*.
Kamis, 13 Juli 2017	1. 4th Weekly Report dengan Prof. Yasuda. 2. Mengimplementasi mekanisme *anonymization* untuk *person* dengan	1. Progress seminggu terakhir tersampaikan ke Prof. Yasuda. Prof. Yasuda menyampaikan bahwa *final report* harus di-

	menggunakan bantuan API *genderize.io*. API digunakan untuk mengekstraksi *gender* (laki-laki atau perempuan) dari *first name*.	*deliver* pada hari Selasa minggu berikutnya dan demo program akan dilakukan pada hari Rabu minggu berikutnya. 2. Fungsi untuk melakukan *anonymization* terhadap *person* yang melakukan *request* ke *genderize.io* API.
Jumat, 14 Juli 2017	1. Mengidentifikasi *temporal phrase* dari *sentence* menggunakan *regular expression*. 2. Melakukan *anonymization* dengan menggeneralisasi *temporal phrase*.	1. Python *script* untuk mengidentifikasi *temporal phrase*. 2. Fungsi untuk melakukan generalisasi *temporal phrase*.

Log Activity – Minggu ke-4

Hari, Tanggal	**Kegiatan**	**Hasil**
Senin, 3 Juli 2017	1. Mendefinisikan *feature* yang akan dilatih pada algoritma CRF. 2. Membuat *script* untuk melakukan *training* model dengan *tools* sklearn-crfsuite.	1. Fungsi yang mengekstraksi *feature* dari *input message.* 2. Python *script* yang dapat melatih model dan menyimpan model.
Selasa, 4 Juli 2017	1. Eksplorasi cara untuk menentukan kedekatan informasi antara dua kata. Dua kata yang dibandingkan yakni kata-kata dalam *social media message* dan kata-kata dalam *user profile*. 2. Ekplorasi beberapa *dataset* yang mungkin. Diputuskan untuk implementasi menggunakan Google. Jumlah kemunculan (*occurrence*) dalam hal ini yakni jumlah halaman web yang muncul saat melakukan pencarian dengan dua kata tersebut.	1. Cara yang dipilih yakni menggunakan *co-occurence metrics* dari *huge dataset.* 2. Python *script* yang memanfaatkan Google Custom Search API untuk menghitung *co-occurence metrics* antara dua kata.
Rabu, 5 Juli 2017	1. Melanjutkan CRF *algorithm script*, membenarkan *bug-bug* yang ditemukan, dan melakukan modularisasi program. 2. Mengeksplorasi WordNet untuk mengukur *sentence similarity* antara *original message* dan *anonymized message*.	1. *Script* telah dapat memprediksi *named entity* dengan format yang benar. 2. Python *script* untuk mengukur *sentence similarity*, namun masih *buggy*.
Kamis, 6 Juli 2017	1. Menyusun metode dalam bentuk *diagram workflow* untuk menentukan frase mana yang benar-benar *private*. Secara umum, dilakukan dalam dua tahap, yakni *co-occurence metrics* (telah dikerjakan) dan *rule-based approached* (belum dikerjakan). *Workflow* ini nantinya akan dibahas bersama Prof. Yasuda pada *review* hari Jumat.	1. Diagram *workflow* / cara kerja program dalam mengidentifikasi beberapa *private phrase* seperti *person*, *location*, *organization*, dan *time*.

Jumat, 7 Juli 2017	1. 4th Weekly Report dengan Prof. Yasuda. 2. Memulai pengerjaan identifikasi *private phrase* dengan *rule-based approach*.	1. Progress seminggu terakhir tersampaikan ke Prof. Yasuda. Mendapatkan *feedback* bahwa untuk *threshold* dari *co-occurence metrics* sebaiknya didefinisikan dalam *range.* Metode yang diajukan untuk untuk mengurangi jumlah *replacement* sudah baik. Namun, lebih bagus lagi jika dirancang sebuah mekanisme untuk menentukan apakah informasi yang tersisa dalam teks dapat digunakan untuk mengidentifikasi seseorang. 2. Fungsi-fungsi Python untuk mengekstrak *candidate phrase* dan menyeleksi *private phrase*.

Log Activity – Minggu ke-3

Hari, Tanggal	**Kegiatan**	**Hasil**
Senin, 26 Juni 2017	1. Membuat *postprocessing module* untuk meng-*construct* kembali kalimat dari struktur nltk.tree.Tree ke struktur kalimat aslinya. 2. Membuat *simple replacer*, yakni mengganti *named entity* dengan kelasnya masing-masing. 3. Menambahkan mekanisme *coreference resolution*, yakni mengidentifikasi seluruh entitas yang sama dan menggantinya dengan frase yang sama pula.	1. Penambahan modul baru, yakni *postprocessing module* pada sistem. 2. Penambahan fungsionalitas baru yakni penggantian *named entity* pada modul *anonymization.* 3. Penambahan fungsionalitas baru yakni *coreference resolution* pada modul *anonymization.*
Selasa, 27 Juni 2017	1. Melakukan evaluasi dari *classifier* yang sudah dibuat dan menganalisis penyebab tingkat akurasi yang rendah dari *classifier.* 2. Mengeksplorasi algoritma-algoritma baru yang meng-*consider feature* dari kata-kata pada satu kalimat. Setelah dipertimbangkan, saya memilih untuk mengeksplorasi lebih jauh algoritma *Conditional Random Field* (CRF).	1. Mengetahui penyebab rendahnya akurasi, yakni *corpus dataset* yang terbatas dan algoritma *training model* yang hanya meng-*consider feature* dari kata yang akan diprediksi saja. 2. Mengetahui prinsip kerja dari algoritma CRF, khususnya *linear chain* CRF.
Rabu, 28 Juni 2017	1. Mengeksplorasi *tools* yang dapat digunakan untuk mengimplementasi algoritma *linear chain* CRF. 2. Melakukan migrasi *corpus* dari *text file* ke satu basis data yang terintegrasi agar mudah untuk menambahkan *corpus* baru (agar *scalable*).	1. Ditentukan *tools* yang dapat digunakan, yakni sklearn-crfsuite versi 0.3. 2. Basis data MySQL berisi data seluruh *corpus* yang digunakan. Basis data menyimpan kata (*word*), *pos tag* dari kata, dan *named entity* dari kata.
Kamis,	1. 3rd Weekly Report dengan Prof. Yasuda	1. Progress seminggu terakhir

29 Juni 2017		tersampaikan ke Prof. Yasuda dan mendapatkan *feedback* bahwa perlu ada mekanisme dan kriteria untuk mengukur seberapa jauh *message* bisa dikatakan sudah *private*. Harapannya, program memiliki kinerja bagus yakni dapat sesedikit mungkin melakukan *replacement* sehingga informasi masih *useful* dan *distance* antara teks asli dan *anonymized text* sedekat mungkin.
Jumat, 30 Juni 2017	1. Melanjutkan proses migrasi *corpus* ke *database* (membutuhkan proses yang lama dan perlu *preprocessing script* yang berbeda-beda untuk setiap *corpus* karena struktur setiap *corpus* berbeda pula). 2. Melakukan eksplorasi penggunaan *tools* sklearn-crfsuite 0.3.	1. Basis data MySQL berisi data seluruh *corpus* yang digunakan. Basis data menyimpan kata (*word*), *pos tag* dari kata, dan *named entity* dari kata. 2. Mengetahui cara implementasi pendefinisian *feature* dan cara melakukan *training* model pada sklearn-crfsuite.

Log Activity – Minggu ke-2

Hari, Tanggal	**Kegiatan**	**Hasil**
Senin, 19 Juni 2017	1. Mempelajari struktur data dari *corpus* yang digunakan. 2. Mempelajari *scikit-learn module* melalui buku "*Hands-On Machine Learning with Scikit-Learn & TensorFlow*" karangan Aurelion Geron Penerbit O'Reilly.	1. Mengetahui bagaimana cara menggunakan dan memanfaatkan *corpus* tersebut. 2. Mengetahui cara membuat *classifier*, melakukan *training* dan *testing* menggunakan *scikit-learn.*
Selasa, 20 Juni 2017	1. Mencoba membuat beberapa *classifier* seperti Naive Bayes dan Perceptron.	1. *Classifier* program untuk *training corpus.*
Rabu, 21 Juni 2017	1. Membuat dokumen *term of reference* (TOR). 2. Melakukan *debugging* terhadap program *classifier.*	1. Dokumen TOR dalam bentuk *hardcopy.* 2. *Bug fixed.*
Kamis, 22 Juni 2017	1. 2nd Weekly Report dengan Prof. Yasuda dan *review* TOR. 2. Melakukan eksperimen *training corpus* dengan *classifier* dan jumlah *sample* data *train* yang berbeda-beda.	1. Progress seminggu terakhir tersampaikan ke Prof. Yasuda dan mendapatkan *feedback* untuk mencoba metode k-*anonymization*. Namun, setelah dieksplorasi, metode tersebut tidak sesuai karena membutuhkan *structured data* sebagai input (semisal *record-record* pada *database*). Sedangkan, *research project* saya menggunakan *unstructured data* (*message*) sebagai input. 1. Model hasil *training.*
Jumat, 23 Juni 2017	1. Melakukan *debugging* kembali terkait masalah-masalah seperti *MemoryError.* 2. Membuat mekanisme *save/load file* model dan melakukan *testing* terhadap sejumlah *input*	1. *Bug fixed* dengan menggunakan *partial fit.* 2. Program *classifier* termodifikasi untuk *save/load*

	message.	model. Mengetahui bahwa *classifier* masih belum akurat dalam mendeteksi kelas yang tepat untuk sebuah *named-entity.*

Log Activity – Minggu ke-1

Hari, Tanggal	**Kegiatan**	**Hasil**
Senin, 12 Juni 2017	1. Tiba di *Kansai International Airport*, Jepang. 2. *Welcome ceremony* dengan Mr. Yoshito Itoh (*School Principal*) di NIT – GC dan pemberian uang beasiswa untuk tiga minggu pertama.	1. - 2. -
Selasa, 13 Juni 2017	1. *Research Introduction* oleh Profesor Tajima (*Computer Network Lab*) dan Profesor Yasuda (*Artificial Intelligence Lab*). 2. Perkenalan diri dengan para siswa yang berada di Prof. Tajima dan Prof. Yasuda Lab.	1. Topik riset terpilih, yakni *Message Anonymization using Trained Named Entity Recognition*. Riset dilakukan di Prof. Yasuda Lab. 2. -
Rabu, 14 Juni 2017	1. Mempelajari *paper* berjudul "*Anonymizing Private Phrases and Detecting Disclosure in Online Social Networks*" yang merupakan disertasi dari Prof. Nguyen Son Hoang Quoc. 2. Kegiatan mingguan *talk cafe*, yakni *foreign students* berinteraksi dengan siswa NIT – GC.	1. Mengetahui variasi metode yang dapat digunakan untuk mendeteksi *named-entity* dan untuk me-*replace named-entity*. 2. -
Kamis, 15 Juni 2017	1. Mempelajari *tools* dan Python *libraries* yang akan digunakan, yakni *Natural Language Toolkit* dari buku karangan Steven Bird, Ewan Klein, dan Edward Loper Penerbit O'Reilly. 2. *1st Weekly Report* dengan Prof. Yasuda. 3. *Welcome party* dengan siswa NIT – GC.	1. Mengetahui cara melakukan *preprocessing* dari *raw text*, cara melakukan *NP chunking*, dan cara melakukan *named entity recognition*. 2. Progress seminggu terakhir tersampaikan ke Prof. Yasuda dan mendapatkan *feedback*. 3. -
Jumat, 16 Juni 2017	1. Mulai membuat program sederhana dalam bahasa Python untuk mendeteksi *named-entity* dengan *natural language toolkit library*.	1. *Source code* program dalam Python yang dapat dilihat di https://github.com/geraldzakwan/TextAnonymizationGNCT .